2 RESEARCH QUESTIONS

- How to integrate ethical values in AI applications development life-cycle using an established set of Software Engineering Methods?
- How to design and develop a process model that can elicit, analyze, and preserve ethical values in AI systems' software design and development?
- How to formally represent, verify, and validate ethical values for AI systems?
- How to identify multiple stakeholders' intended role in an Artificial intelligence software development environment where the process is continuously planned, developed, tested, integrated, and delivered?
- How to design and develop implementation techniques that stakeholders could use to evaluate and maintain responsible development of artificially intelligent machines?

3 DEVELOPMENT PROCESS

The process of identifying, integrating, designing, implementing, evaluating, and maintaining ethical values in the design of AI systems requires its methods to be systematically structured with the development life-cycle, participation support, and trustworthiness. Software development methodologies and best engineering practices to practically guide the system engineering process with ethical principles by stakeholders' active involvement can provide a useful quality management framework for solving challenges related to the development of responsible AI systems. Implementation and standardization of engineering practices are essential for developing high-quality and robust AI systems adhering to ethical values and societal norms. Therefore, it is essential to evaluate how current methodologies and their applications can support these issues. The standards of trustworthiness require soft guidelines to be translated into formal specifications using a well-established set of SE techniques. [1].

AI systems are also artifacts designed and developed by humans that require an established set of engineering practices. While different approaches, ontologies, taxonomies may be required for different AI application domains, trustworthiness criteria can be satisfied by applying established SE methods. A framework that provides a mechanism to integrates ethical values as exact requirements specification in the system engineering process by involving all stakeholders. Similar to how a "traditional" software life cycle starts with the requirements engineering phase, building a concrete policy starts with the ethical-requirements management phase. Then, it moves into assigning new "or mapping existing" governance mechanisms to satisfy the said requirements. Once this assignment is completed, compliance can be determined in the next step. This phase aims to set concrete requirements, each of them a moral value, on not only what high-level ethical principles the system needs to adhere. This framework will provide methods for fostering ethical, legal, social, economic, and cultural values by converting them into functional specifications that will inform the system's intent and operational interpretation transparently.

In addition to developing well-established SE life-cycle support mechanisms for AI systems and assessing engineering processes maturity, it is vital to ensure that AI system engineering design choices also support Socio-Technical Systems (STS) engineering approaches. Such approaches to system development lead to more acceptable systems to end-users and deliver better value to stakeholders through implementing and re-evaluating ethical values in the initial design phase during system design. Maintaining a continuous and transparent reporting of the development process decisions and choices can be interpreted, verified, validated, and, if necessary, modified.

4 CONTRIBUTION AND FUTURE WORK

As the first step in this research, different aspects of existing AI engineering methods are explored and analyzed for different AI application domains such as *socio-technical systems, multi-agent systems, agent-based social simulations, architectures for embodied agents,* and *machine learning*. Selected methods are *OperA, EI/EIDE, JaCaMo, ABSS, Behaviour Tree, Goal-Oriented Action Planning, Behaviour-Oriented Design, FAtiMA*, and *SEMMA*. Existing AI development methods used in various AI applications domains do not wholly support SDLC. Furthermore, none of the existing AI development methods wholly support participation support and trustworthiness in the development process. Another survey related to ethical guidelines for trustworthy AI was performed. Our survey exposed various shortcomings in existing development methods and how lack of structured SE life-cycles approaches can affect the development of high-quality, transparent, and trustworthy AI systems. On the other hand, the survey also identifies some key findings for future methodology and how various operational elements and metrics from a well-established software development process can lead stakeholders and their development activities towards trustworthiness.

This PhD research aims to engineer a design methodological framework to elicit and analyze ethical values to be aligned with system goals and provide a mechanism for explicit interpretation, formal representation, verification, and validation of values. This framework will also specify governance mechanisms to ensure openness and support mechanisms for multiple stakeholders' participation. In the next step, ethical values will be systematically analyzed and mapped with SDLC, making it accessible for multiple stakeholders to do the ethical assessment of key values required for the desired system. Later the values will formally be verified and validated using formal methods.

5 ACKNOWLEDGEMENTS

This work is supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] Pierre Bourque and Richard E. Fairley (Eds.). 2014. *SWEBOK: Guide to the Software Engineering Body of Knowledge* (version 3.0 ed.). IEEE Computer Society, Los Alamitos, CA. http://www.swebok.org/
[2] Virginia Dignum. 2017. Responsible Artificial Intelligence: Designing AI for Human Values. *ICT Discoveries* 1 (2017), 1–8.
[3] Virginia Dignum. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way.* Springer International Publishing.
[4] A Theodorou. 2019. *AI Governance Through A Transparency Lens.* Ph.D. Dissertation. University of Bath.
[5] Andreas Theodorou and Virginia Dignum. 2020. Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence* 2, 1 (2020), 10–12.